

https://doi.org/10.51885/3134-8025_IICS_2026_3_2
SRSTI 81.93.29

ENHANCING INTRUSION DETECTION SYSTEMS WITH AI-BASED NETWORK TRAFFIC ANALYSIS: A FOCUS ON ANOMALY DETECTION

ЖАСАНДЫ ИНТЕЛЛЕКТКЕ НЕГІЗДЕЛГЕН ЖЕЛЛІК ТРАФИКТІ ТАЛДАУ АРҚЫЛЫ ИНТРУЗИЯНЫ АНЫҚТАУ ЖҮЙЕЛЕРІН ЖЕТІЛДІРУ: АУЫТҚУЛАРДЫ АНЫҚТАУ

УСОВЕРШЕНСТВОВАНИЕ СИСТЕМ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ С ПОМОЩЬЮ АНАЛИЗА СЕТЕВОГО ТРАФИКА НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ФОКУС НА ОБНАРУЖЕНИИ АНОМАЛИЙ

G. Ziyatbekova^{1,2*}, G. Kydyrbayeva³, G. Rahymzhanova⁴, G. Shametova¹,
B. Moldakalykova⁵, N. Bilimzhankyzy⁵

¹Al-Farabi Kazakh National University, Almaty, Kazakhstan

²Almaty Technological University, Almaty, Kazakhstan

³I. Zhansugirov Jetisu University, Taldykorgan, Kazakhstan

⁴Educational and training center "Daryn", Taldykorgan, Kazakhstan

⁵Turan University, Almaty, Kazakhstan

*Corresponding author: Gulzat Ziyatbekova, e-mail: ziyatbekova1@gmail.com

Keywords:

NGFW, IDS/IPS, data leakage
regression, machine learning,
network security, anomaly
detection, cybersecurity,
Random Forest.

ABSTRACT

In an era of sophisticated cyber threats and the growing complexity of network infrastructures, intrusion detection systems (IDS) play a critical role in ensuring the security of digital assets. Traditional IDS methods based on predefined rules and signatures are becoming less effective against sophisticated attacks. The study covers the stages of data collection, preprocessing, model training, and evaluation. This study examines current security solutions, such as next-generation firewalls (NGFWs) and intrusion detection and prevention systems (IDS/IPS). In the experimental study, linear regression, Ridge regression, Lasso regression and Random Forest models were used. Particular attention is paid to the methodological problem of data leakage. Special attention was paid to the methodological problem of data leakage, and a leakage-free modeling scenario based only on previously known variables was created. Taking into account the pronounced right-hand asymmetry and heavy-tailed distribution of the dependent variable, the effect of using the log1p transformation on the generalization ability of the models was investigated. The obtained results demonstrate the advantages of regularized linear models. The use of log1p transformation in combination with Ridge regression allowed us to improve the quality of forecasting, reduce the variance of errors and increase the value of the determination coefficient R^2 . Additional validation based on time-based data partitioning confirmed the methodological correctness of the





conclusions drawn. The Random Forest model did not show any improvement in forecasting accuracy, but was used to analyze the structure of factors influencing the engagement rate. In addition, issues such as model interpretability, false positives, and the need for continuous learning are addressed. An intelligent approach based on event log analysis and machine learning methods is proposed to identify signs of potential information leaks.

Түйінді сөздер:

NGFW, IDS/IPS,
деректердің ағып кетуі,
регрессия, машиналық
оқыту, желілік
қауіпсіздік, аномалияны
анықтау,
киберқауіпсіздік, Random
Forest.

ТҮЙІНДЕМЕ

Киберқауіптердің күрделілігі және желілік инфрақұрылымдардың күрделенуі дәуірінде бұзушылықтарды анықтау жүйелері (IDS) цифрлық активтердің қауіпсіздігін қамтамасыз етуде маңызды рөл атқарады. Алдын ала анықталған ережелер мен қолтаңбаларға негізделген дәстүрлі IDS әдістері күрделі шабуылдарға қарсы тиімділігін жоғалтуда. Зерттеу деректерді жинау, алдын ала өңдеу, модельді оқыту және бағалау кезеңдерін қамтиды. Бұл зерттеу келесі буын брандмауэрлері (NGFW) және енуді анықтау және алдын алу жүйелері (IDS/IPS) сияқты қазіргі қауіпсіздік шешімдерін қарастырады. Тәжірибелік зерттеуде сызықтық регрессия, Ridge, Lasso регрессиясы және Random Forest модельдері қолданылды. Деректердің ағып кетуінің (data leakage) әдіснамалық мәселесіне ерекше назар аударылып, тек алдын ала белгілі болатын айнымалылар негізінде деректердің ағып кетуі жойылған (leakage-free) модельдеу сценарийі қалыптастырылды. Сонымен қатар тәуелді айнымалының айқын оңға асимметриялы және heavy-tailed таралғанын ескере отырып, log1p түрлендіруінің модельдердің жалпылау қабілетіне әсері зерттелді. Алынған нәтижелер реттелген сызықтық модельдердің артықшылықтарын көрсетеді. log1p түрлендіруін Ridge регрессиясымен бірге қолдану болжау сапасын жақсартуға, қателіктердің дисперсиясын азайтуға және R^2 детерминация коэффициентінің мәнін арттыруға мүмкіндік берді. Уақытқа негізделген деректерді бөлуге бағытталған қосымша валидация жасалған қорытындылардың әдіснамалық дұрыстығын растады. Кездейсоқ Random Forest моделі болжам дәлдігінде ешқандай жақсару көрсеткен жоқ, бірақ қатысу деңгейіне әсер ететін факторлардың құрылымын талдау үшін пайдаланылды. Сонымен қатар, модельді түсіндіру, жалған оң нәтижелер және үздіксіз оқыту қажеттілігі сияқты мәселелер қарастырылады. Ақпараттың ағып кетуінің ықтимал белгілерін анықтау үшін оқиғалар журналын талдау және машиналық оқыту әдістеріне негізделген интеллектуалды тәсіл ұсынылады.

Ключевые слова:

NGFW, IDS/IPS, утечка
данных, регрессия,
машинное обучение, сетевая
безопасность, обнаружение
аномалий,
кибербезопасность, Random
Forest.

АННОТАЦИЯ

В эпоху сложных киберугроз и растущей сложности сетевых инфраструктур системы обнаружения вторжений (IDS) играют критически важную роль в обеспечении безопасности цифровых активов. Традиционные методы IDS, основанные на заранее определенных правилах и сигнатурах, становятся менее эффективными против сложных атак. Исследование охватывает этапы сбора данных, предобработки, обучения модели и оценки ее работы. В настоящем исследовании рассматриваются актуальные средства защиты, такие как межсетевые экраны нового поколения (Next-Generation Firewall, NGFW) и системы обнаружения и предотвращения вторжений (IDS/IPS). В рамках экспериментального исследования применены модели линейной регрессии,

Ridge-перпессии, Lasso-перпессии и Random Forest. Особое внимание уделено методологической проблеме утечки данных (data leakage), и был создан сценарий моделирования, исключающий утечку данных (leakage-free), основанный только на заранее известных переменных. С учётом выраженной правосторонней асимметрии и heavy-tailed распределения зависимой переменной было исследовано влияние применения преобразования $\log_{1p}$ на обобщающую способность моделей. Полученные результаты демонстрируют преимущества упорядоченных линейных моделей. Использование $\log_{1p}$ -преобразования в сочетании с Ridge-регрессией позволило нам улучшить качество прогнозов, уменьшить дисперсию ошибок и повысить значение коэффициента детерминации R^2 . Дополнительная проверка, сфокусированная на дезагрегации данных по времени, подтвердила методологическую обоснованность сделанных выводов. Модель Random Forest не показала улучшения точности прогнозирования, но использовалась для анализа структуры факторов, влияющих на показатели участия. Кроме того, рассматриваются такие вопросы, как интерпретация модели, ложные срабатывания и необходимость непрерывного обучения. Предлагается интеллектуальный подход, основанный на анализе журналов событий и методах машинного обучения, для выявления потенциальных признаков утечки информации.

INTRODUCTION

Modern cyber threats are characterized by a high degree of complexity and variability, which makes the task of ensuring the protection of corporate networks from the leakage of confidential information particularly urgent. Traditional approaches to information security are becoming less effective when faced with attacks based on atypical user behavior and the use of hidden communication channels. In this regard, the role of next-generation firewalls (NGFW) and intrusion detection and prevention systems (IDS/IPS) is increasing (Tsai, CF, Hsu, YF, Lin, CY, & Lin, WY, 2009).

This study aims to explore the possibilities of integrating NGFW and IDS/IPS with machine learning technologies to implement intelligent network traffic analysis and identify potential data leakage channels (Hamid, Y., Sugumaran, M., & Balasaraswathi, VR, 2016). As part of the practical part, a module was developed in the Python programming language (Kumar, I., Bhatt, C., Patel, R., & others, 2020), based on the Random Forest algorithm, which provides automatic classification of threats. The module was tested on various attack scenarios to assess its effectiveness and applicability in real-world conditions.

The purpose of the study is to conduct a comparative analysis of the effectiveness of firewalls, to develop an intelligent system for protecting a corporate network from data leaks based on NGFW, IDS/IPS and machine learning methods. The study analyzes the main factors influencing the effectiveness of creating an intelligent system and conducts a comparative analysis of various machine learning models for forecasting.

The main scientific and methodological contribution of this work is the creation of a data leakage-preventing assessment scenario for predicting the protection of corporate networks from information leaks using only available indicators, as well as additional verification of the obtained results. Furthermore, the effect of $\log_{1p}$ transformation on the generalizability of regularized linear models (Ridge/Lasso) in the case of a right-skewed, heavy-tailed distribution of the dependent variable is empirically demonstrated (Dua, M., & Kunal., 2019).

MATERIALS AND METHODS

The dataset used in the study defines the characteristics and quantitative indicators of data in corporate networks. The data was collected from open sources. The dataset consists of 500 rows (Theodoridi P.K., & Gkikas D.C., 2025).

Most of our machines 4 models of coolers. The first is the linear regression model, which is the basic model for intrusion detection. Linear regression coefficients allow us to quantify the influence of variables on the predicted indicator. This is important for interpreting the modeling results. However, the main drawback of the simple linear regression model is that it cannot model complex nonlinear relationships between variables. In addition, linear regression is sensitive to correlations between variables.

Limitations of linear models can be overcome with the help of regularization methods, such as Ridge and Lasso regression. Ridge regression is based on the principle of L2 regularization and improves the stability of the model by bringing all coefficients closer to zero. This method is particularly effective when there is a high correlation between variables. The Lasso regression method allows some coefficients to be completely set to zero using L1 regularization. It is this property that allows the Lasso model to be used as a tool for variable selection. As a result, the model becomes more compact and the level of interpretability increases (Andariesta D.T., & Wasesa M., 2023).

Of the nonlinear models, the Random Forest model was used. It is based on decision trees and combines the results of multiple trees to improve the robustness of the model. This model is robust to noise and is able to estimate the significance of variables. For the Random Forest model, key hyperparameters (n_estimators, max_depth, min_samples_leaf) were tested at several values, and the configuration that yielded stable results was selected. The purpose of this comparative analysis was not to identify the most complex model, but to evaluate the additional benefits of model complexity in a leakage-free scenario (Biau G., & Scornet E., 2016; Davies S.R., Wells R., Zollo F., Roche J., 2024).

RESULTS AND DISCUSSION

A common problem in intrusion detection systems based on network traffic analysis is methodological errors in model construction, including data leakage (Liu, H., & Lang, B., 2019). Including variables such as reach, impressions, and consumption, which are used to identify anomalies in network traffic analysis, allows the predictive model to provide better results (Kapoor S., & Narayanan A., 2022).

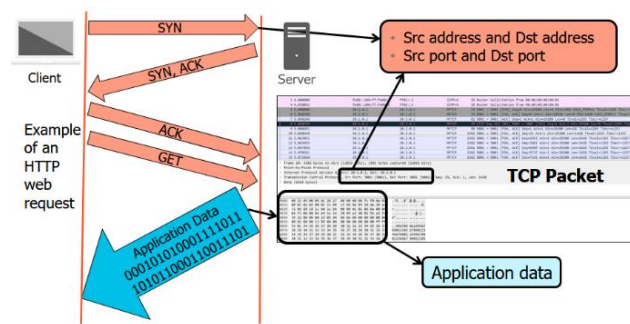


Figure 1. Threat detection process

The integration of modern firewalls with other elements of the information security architecture, including IDS/IPS systems, can significantly improve the complexity and speed of response to cyber threats (Figure 1). Intrusion detection and prevention systems (IDS/IPS) occupy a central place in the mechanisms of network exchange analysis and operate in two basic modes:

monitoring (IDS) and intervention (IPS). The first type aims to detect suspicious activity without intervention, while the second not only diagnoses threats but also initiates actions to immediately suppress them (Table 1).

Table 1. Differences between IDS and IPS

Characteristic	IDS	IPS
Reaction	Detection and notification	Preventing and blocking threats
Actions	Passive monitoring	Active intervention
Usage	Log analysis, investigation	Real-time preventative protection
<i>Note – compiled by the authors</i>		

For the purpose of structural comparison of the characteristics of different types of IDS/IPS, it is advisable to consider the summary table (Table 2), which shows the key parameters and features of network, host and hybrid systems:

Table 2. IDS/IPS Types

System type	Advantages	Restrictions	Examples of threats
Network IDS/IPS (NIDS/NIPS)	Effective control of network traffic at the network perimeter	They do not protect actions on hosts and are vulnerable to encryption	DDoS, SQL injection, flooding
Host-based IDS/IPS (HIDS/HIPS)	Deep control of events on endpoints	They don't analyze network traffic and are difficult to scale	Viruses, Trojans, internal leaks
Hybrid IDS/IPS	Comprehensive protection, wide coverage	High resource requirements, difficult to administer	APT, advanced threats, combined attacks
<i>Note – compiled by the authors</i>			

Thus, understanding the characteristics of different types of IDS/IPS allows you to choose the most appropriate approach to protecting your corporate network. In today's environment, hybrid systems that combine the capabilities of NIDS/NIPS and HIDS/HIPS, as well as those that use machine learning to adapt to new types of threats, are considered the most effective (Thakkar, A., & Lohiya, R., 2021).

This study presents a comprehensive assessment of the functionality of next-generation firewalls (NGFW) in the context of countering data leaks in corporate networks using modern analytical methods. *The scientific novelty of the work* lies in the development and integration of the author's NGFW log analysis module, based on a behavioral approach to identify potential information leakage incidents. Unlike classic signature-based systems, the proposed solution uses machine learning algorithms and heuristic techniques to detect anomalies in network traffic. The study conducted an empirical comparison of the effectiveness of traditional NGFWs with intelligent leak detection systems based on machine learning methods. The Random Forest classification algorithm was applied, which demonstrated the superiority of behavioral analysis in identifying non-standard and previously unknown data compromise scenarios. Particular attention was paid to the creation of an experimental setup, including a virtualized infrastructure with a deployed NGFW and simulated data leak scenarios. This ensured the reproduction of realistic operating conditions of the corporate network and allowed for a full verification of the developed methods.

The author also proposed a classification of signs characteristic of various types of leaks, including the transmission of encrypted information through non-standard ports, the use of tunneling protocols, and suspicious network activity during non-working hours. The generated database of anomalous patterns can be effectively used for automated configuration of monitoring systems in real organizational infrastructures.

The practical value of the research lies in the possibility of integrating the created software module into corporate information security systems, as well as in the use of machine learning methods to improve the level of protection against threats associated with the unauthorized transfer of confidential data outside the internal network.

With the spread of cloud platforms, virtualization and mobile devices, the topology of corporate networks has undergone fundamental transformations. The emergence of distributed computing environments, the concept of Bring Your Own Device (BYOD), and the use of SaaS services have significantly expanded potential attack vectors. In parallel with the evolution of threats, the requirements for protection mechanisms have also evolved (Sudhir, V., 2020).

The answer to these challenges was the Next Generation Firewall (NGFW) - comprehensive solutions that integrate the functions of classic firewalls, intrusion prevention systems (IPS), application control and traffic intelligence. NGFWs provide deep packet inspection (Deep Packet Inspection, DPI), application recognition and control, encrypted traffic processing (including TLS 1.3), user-level filtering, and integration with external threat intelligence systems.

Thus, NGFWs represent a logical stage in the evolution of cyber defense tools, providing multi-layered protection – from the network level to the application, user and content levels. This is especially important in the era of advanced targeted attacks (APTs) used by both cybercriminal groups and nation-state actors. The implementation of artificial intelligence and machine learning enables NGFW to identify previously unknown threats through behavioral analysis and dynamic adaptation to the changing cyber risk landscape (Bowe, Zh., Xiaoliang, W., Ru Xie, Chuncao Li, Huazheng, Zh., Frank, J., 2023).

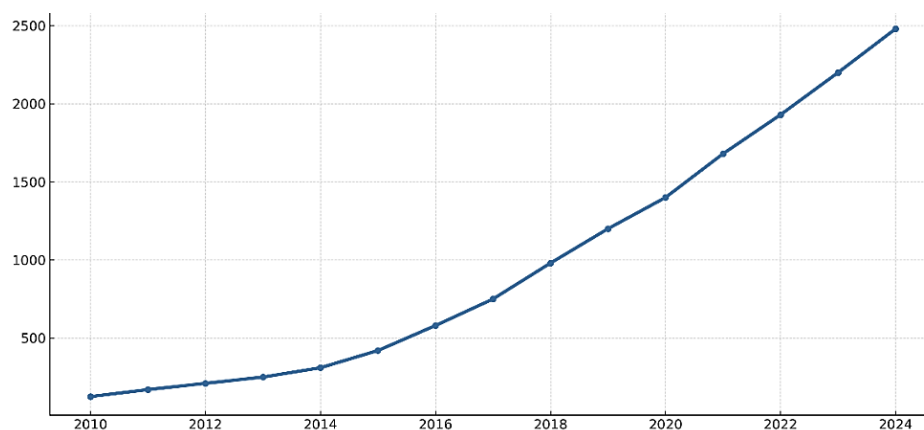


Figure 2. Growth in the number of cyberattacks and information security incidents (2010–2024)

Note – compiled by the authors

The graph (Figure 2) shows the exponential growth of reported cybersecurity incidents (the y-axis shows the number of incidents (thousands)) over the past 15 years, according to IBM X-Force and ENISA. There is a clear correlation between the evolution of threats and the need to move from classic firewalls to NGFWs. A particularly sharp increase has been observed since 2020, which is associated with the mass transition to remote work, the development of cloud technologies, and the activation of APT groups.

Next-Generation Firewall Architecture (NGFW). The typical architecture of an NGFW solution involves step-by-step traffic processing: session analysis, application identification, verification of compliance with known threat signatures, behavioral analysis, and decision-making on whether to allow or block a network connection. A similar concept is implemented, for example, in the products of Palo Alto Networks.

When implementing security systems in large corporations, it is necessary to take into account not only the threats, but also the architectural features of the corporate network. Let's consider a hypothetical case of integrating a security system into an international corporation with a distributed infrastructure, including several data centers and branches around the world. The main threats to this organization include DDoS attacks, phishing, and risks of confidential data leakage (Ashrafi, N., Kuilboer, J.P., Joshi, C., & Ran, W., 2022).

Comparative analysis of popular NGFW solutions. Let's compare the capabilities and functional features of the most popular NGFWs available on the market. Table 3 shows the key evaluation parameters.

Table 3. Comparison of NGFW solutions by key parameters

Parameter	Palo Alto Networks	pfSense	Mikrotik
APT detection	Yes	Partially	Partially
TLS 1.3 support	Yes	Yes	Partially
Built-in sandbox analysis	Yes (WildFire)	Yes	Via AMP
ML/AI support	Yes	Yes	No
Built-in DLP	Yes	Through modules	No
Integration with TI	Automatic	Partially	No
TCO for 5 years (conditional)	Average	Short	Short
<i>Note – compiled by the authors</i>			

Analyzing the data provided (Table 3), it can be noted that Palo Alto Networks solutions demonstrate the highest level of integration of key security functions. These include full support for TLS 1.3, the use of behavioral analysis using machine learning and artificial intelligence technologies, built-in protection mechanisms against sophisticated APT attacks, and effective implementation of sandbox analysis through the WildFire platform. This is why these products are often chosen to protect the infrastructure of large corporations and critical information systems. Automatic integration with global threat intelligence platforms allows Palo Alto's NGFW to adapt to emerging threats without requiring administrator intervention (Gruschka, N., Jensen, M., & Iacono, L., 2023).

Therefore, none of the solutions considered is universally applicable to all situations. The choice of NGFW should be based on the specific requirements of the organization, taking into account factors such as:

- criticality of protected resources;
- acceptable delays in traffic analysis;
- budget for implementation and operation;
- the need for integration with existing monitoring and analysis systems.

The comparative analysis carried out clearly demonstrates the importance of a systematic approach to the selection of NGFW, based on a three-level methodology presented in Figure 2, which takes into account technical characteristics, security functionality and economic indicators.

So, it can be noted that modern NGFWs represent a qualitative step forward compared to classic firewalls. They are capable of effectively countering multi-layered cyber threats, offering an advanced set of features: deep packet inspection (DPI), application-level filtering, integration

with threat intelligence, behavioral analysis and machine learning, as well as advanced sandbox malware analysis.

The NGFWs reviewed – Palo Alto Networks, Fortinet FortiGate, Cisco Firepower, and Check Point – each have their own strengths and weaknesses, highlighting the need for a thorough and comprehensive approach when choosing a security system.

NGFWs are therefore a critical component of a modern corporate network cybersecurity strategy, enabling adaptation to a variety of threat scenarios and using intelligent methods to improve protection effectiveness. The main task is to classify threats based on log data using the Random Forest algorithm, one of the most effective methods for working with large volumes of information and building accurate classifiers.

Sample code for training a Random Forest model. To begin, let's create an IDS/IPS model in Python using the scikit-learn library. Code example:

```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
dataset = pd.read_csv('network_traffic_data.csv')
features = dataset.drop('attack_type', axis=1) # All parameters except class label
target = dataset['attack_type'] # Type of threat identified
X_train, X_test, y_train, y_test = train_test_split(
    features, target, test_size=0.2, random_state=42
)
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
predictions = model.predict(X_test)
print(classification_report(y_test, predictions))
```

Before starting training, it is necessary to split the initial sample into training and test segments. This procedure is performed using the `train_test_split` function, which ensures a random and proportional distribution of the data. This separation makes it possible to objectively test the model on previously unused examples, thereby allowing us to evaluate its ability to generalize.

After completing the training phase, the model is tested on test data, and its accuracy is analyzed using the `classification_report` function from the same library.

DDoS attack. Distributed Denial of Service (DDoS) attacks are designed to deplete the network or computing resources of a target system in order to temporarily disable it. Such actions are especially dangerous in a corporate environment, as temporary service downtime can lead to direct losses and damage to the organization's reputation.

As part of the experiment, a DDoS attack was simulated using the `hping3` utility running in Kali Linux. This tool allows you to generate arbitrary network packets and implement scenarios typical for automated botnets. The test attack used the SYN flood technique, one of the most well-known types of DDoS, which involves sending a large number of TCP SYN packets but does not complete the three-way handshake process. This load leads to the exhaustion of the resources of the server responsible for processing incoming connections and disruption of its stable operation (Figure 3).

```
(root@kali)~[/home/kali]
# hping3 -S --flood -V -p 80 172.16.20.251
using eth0, addr: 172.16.20.22, MTU: 1500
HPING 172.16.20.251 (eth0 172.16.20.251): S set, 40 headers + 0 data bytes
hping in flood mode, no replies will be shown
```

Figure 3. Carrying out a DDoS attack using the hping method

Note – compiled by the authors

Detailed Log View													
General				Source				Destination					
Session ID 0				Source User				Destination User					
Action random-drop				Source 0.0.0.0				Destination 0.0.0.0					
Host ID				Source DAG				Destination DAG					
Application not-applicable				Country 0.0.0.0-0.255.255.255				Country 0.0.0.0-0.255.255.255					
Rule				Port 0				Port 0					
Rule UUID				Zone LAN				Zone LAN					
Device SN				Interface				Interface					
IP Protocol tcp				X-Forwarded-For IP									
Log Action								Flags					
Generated Time 2025/04/20 02:24:49													
PCAP	RECEIVE TIME	TYPE	APPLICAT...	ACTION	RULE	RULE UUID	BY...	SEVERI...	CATEG...	URL CATEG... LIST	VERDI...	URL	FILE NAME
	2025/04/20 02:26:57	flood	not-applicable	random-drop				critical	any				
	2025/04/20 02:26:46	flood	not-applicable	random-drop				critical	any				
	2025/04/20 02:26:38	flood	not-applicable	random-drop				critical	any				

Figure 4. Palo Alto NGFW response to a DDoS attack

Note – compiled by the authors

A DDoS attack simulation confirmed the high effectiveness of Palo Alto's NGFW in identifying and blocking network threats. Using a combined approach that includes signature analysis and behavioral analytics, the system promptly recognized anomalous activity, classified it as a TCP flood attack, and activated protective mechanisms. All events were recorded in detail in a threat log, which ensured transparency of the incident and the possibility of subsequent analysis (Figure 4).

The recorded data became part of the training set for the next stage – machine threat analysis using the Random Forest algorithm. Thus, the experiment confirmed that the integration of NGFW into the IDS/IPS system architecture allows not only to repel attacks in real time, but also to increase the accuracy of intelligent models due to reliable labels and attack indicators.

The results obtained demonstrate the practical value of the proposed solution for comprehensive protection of a corporate network from denial-of-service attacks and similar threats.

This multi-layered approach provides a high degree of reliability and flexibility, allowing corporate networks to effectively counter a wide range of threats, protect data from leaks, and reduce business risks (Biau G., Scornet E., 2016).

This code snippet implements important steps in preparing data for training a machine learning model aimed at classifying threats detected in firewall logs (Figure 5). For each category, the fit transform method is applied, which replaces text values with numeric ones, making the data ready for processing by the model.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix

# === 1. Загрузка и разметка данных ===
df = pd.read_csv("threat.csv")
df['Type'] = df['Type'].astype(str).str.lower()
df['label'] = df['Type'].apply(lambda x: 1 if x in ['threat', 'wildfire'] else 0)

# === 2. Предобработка ===
df = df.dropna(subset=['Application', 'Source Zone', 'Destination Zone', 'Action', 'Severity', 'Threat/Content Type'])

features = ['Application', 'Source Zone', 'Destination Zone', 'Action', 'Severity', 'Threat/Content Type']
X = df[features]
y = df['label']
```

Figure 5. Importing libraries and datasets for analysis using the Random Forest algorithm

Note – compiled by the authors

```
1 Domain,Receive Time,Serial #,Type,Threat/Content Type,Config Version,Generate Time,Source address,Destination address,NAT Source IP,NAT Destination IP,Rule,Source User,
2 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
3 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
4 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
5 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
6 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
7 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
8 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
9 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
10 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
11 1,2025/04/19 19:11:31,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:31,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
12 1,2025/04/19 19:11:15,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:15,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
13 1,2025/04/19 19:11:15,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:15,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
14 1,2025/04/19 19:11:15,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:15,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
15 1,2025/04/19 19:11:15,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:15,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
16 1,2025/04/19 19:11:15,007951000569758,THREAT,vulnerability,2818,2025/04/19 19:11:15,172.16.20.22,192.168.40.15,,LAN-TO-WAN,,web-browsing,vssys1,LAN,DH2,ethernet1/2,ethe
```

Figure 6. Datasets for analysis

Note – compiled by the authors

The data is then split into training and test sets using the `train_test_split` function: 80% of the data is used to train the model, and 20% is used to test its performance on new data. This allows us to evaluate the generalization ability of the model (Figure 6).

Three numerical metrics are used to objectively assess the quality of machine learning models.

MAE (Mean Absolute Error) shows the average value of the absolute values of the forecast errors. This indicator allows us to estimate the average level of model error in real units and is easy to interpret. Because the mean absolute error (MAE) is less sensitive to large deviations, it is often used when evaluating stable models.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

RMSE (Root Mean Squared Error) is the square root of MSE, which allows us to express the forecast error in its original unit of measurement. The root mean square error (RMSE) is very sensitive to large errors and is often used to compare the forecasting accuracy of models.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

The coefficient of determination R^2 shows what proportion of the data variation is explained by the model. This metric ranges from 0 to 1, with a higher value indicating a better fit between the model and the data. R^2 allows you to assess the overall quality of a model, but it is recommended to use it in conjunction with other error metrics.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (3)$$

n – number of observations (total number of elements in the sample);

y_i – (actual) value of the dependent variable;

$\tilde{y}_i$ – the value of the dependent variable predicted by the model;
 $\bar{y}$ – the average value of all actual observations.

The variable "Number of lifetime engaged users" was selected as the target variable. This indicator reflects the number of unique users. During the initial analysis, it was discovered that there were a number of variables present in the data set. Using these variables in a model introduces data leakage, meaning the model may produce artificially inflated results by using information that may not be available in the future. Therefore, these variables were completely excluded from the experimental study.

The dataset contains empty values for some variables, which were filled in two ways:

- ⊙ numeric variables with median values;
- ⊙ Categorical variables were filled with the most frequently occurring values.

Categorical variables were converted to numerical format using One-Hot Encoding method for use in machine learning models. This method is based on using only the values 1 and 0. While the original dataset contained 19 variables, the One-Hot Encoding method resulted in a total of 56 variables. For example, if the original dataset is described by only one variable, then using One-Hot Encoding this variable is multiplied by 7 variables. This approach allows models to treat each value of categorical variables as a separate feature (Davies S.R., Wells R., Zollo F., Roche J., 2024; Eslami S.P., Ghasemaghaei M., Hassanein K., 2022).

The preprocessed dataset was divided into two parts for training and evaluating models: training set – 80%. The distribution process was carried out using a constant initial value of the random number generator to control randomness, which ensures the reproducibility of the experimental results.

All calculations were performed in the Python programming language using the NumPy, pandas, scikit-learn and matplotlib libraries. Artificial intelligence tools were used in the text of the article for scientific structuring of the research results, their grammatical and stylistic unification, and for translation of abstracts.

The distribution of the target variable Lifetime Engaged Users is clearly right-skewed and heavy-tailed. This can be seen from the distribution graph of the target variable in Figure 7. Figure 8 shows the distribution of the target variable after log1p transformation and the boxplot representation. After log1p transformation, the distribution becomes closer to symmetry and the influence of outliers is significantly reduced.

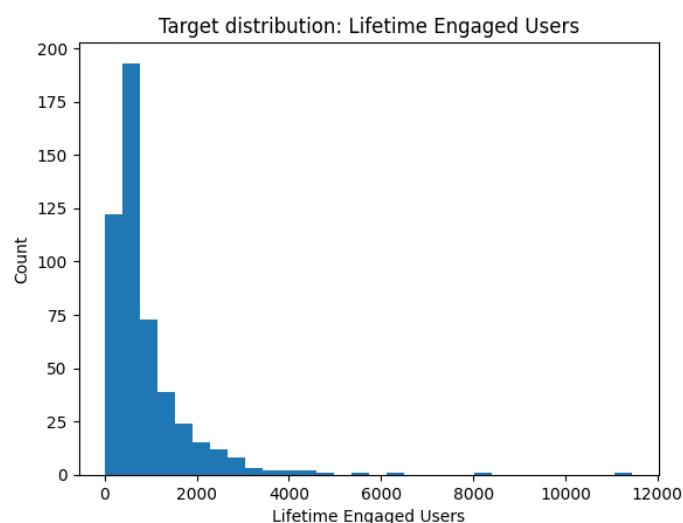


Figure 7. Initial distribution of the Lifetime Engaged Users variable

Note – compiled by the authors

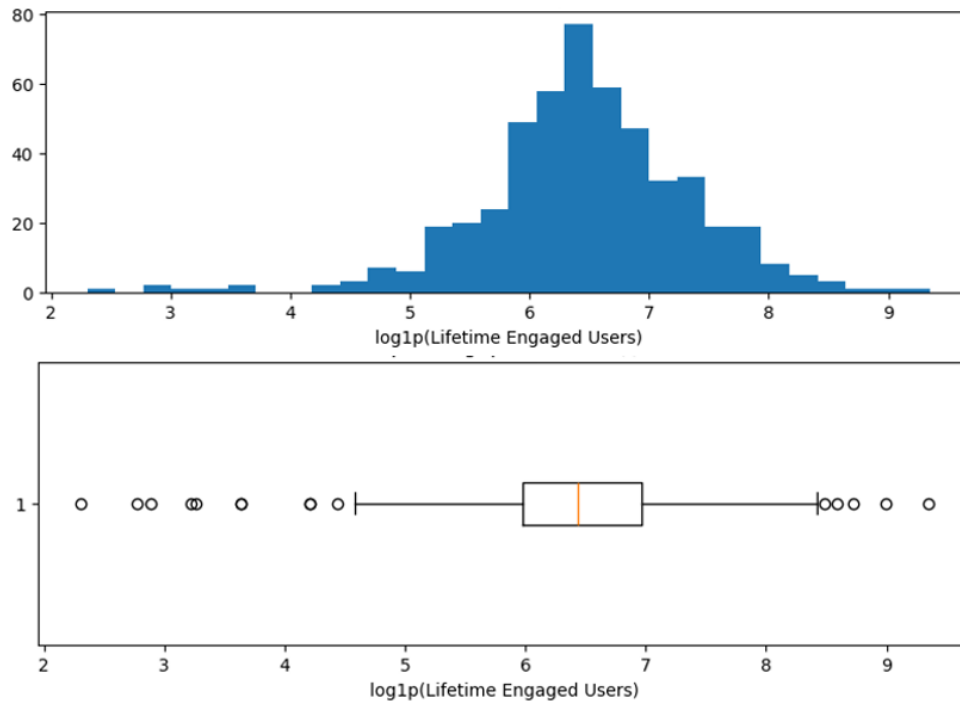


Figure 8. Distribution and boxplot representation of Lifetime Engaged Users after log1p transformation

Note – compiled by the authors

An asymmetric distribution of the target variable can enhance the influence of outliers and increase the error of classical regression models. Therefore, we applied a logarithmic transformation to the target variable, approximating it with a symmetric distribution. The results of descriptive statistics confirmed a high value of the asymmetry coefficient (4.52), a large kurtosis index (34.11) and the proportion of outliers identified by the interquartile range rule of approximately 8.8%. Such a distribution may negatively affect the predictive power of classical linear regression models and therefore indicates the need for a transformation of the target variable.

At first, DummyRegressor (averaging strategy) was used as the base model. Since the target variable has heavy tails, median prediction along with the mean strategy could also be considered as a base case, but since the main goal is to compare models under the same leaky conditions, the standard DummyRegressor(mean) was taken as the base case. The baseline model showed $RMSE \approx 758$ and $R^2 \approx -0.06$, indicating insufficient predictive ability of the model.

Ridge regression trained in the leak-free scenario yielded $RMSE \approx 731$ and $R^2 \approx 0.02$ when using the target variable at its original scale. This result suggests that known factors explain this indicator only to a limited extent.

During the first experiment, it was observed that the forecasting quality was artificially inflated when variables were included in the model. In this case, the model does not predict the future level, but only recalculates the result. In this case, the model does not predict the future level, but only recalculates the result.

Given the obvious skewness and heavy-tailed distribution of the target variable, a log1p transformation was applied to the Lifetime User Engagement metric. As a result, the forecasting performance of ridge regression improved significantly (Figure 9): the RMSE value decreased to approximately 630, and the R^2 value reached 0.27.

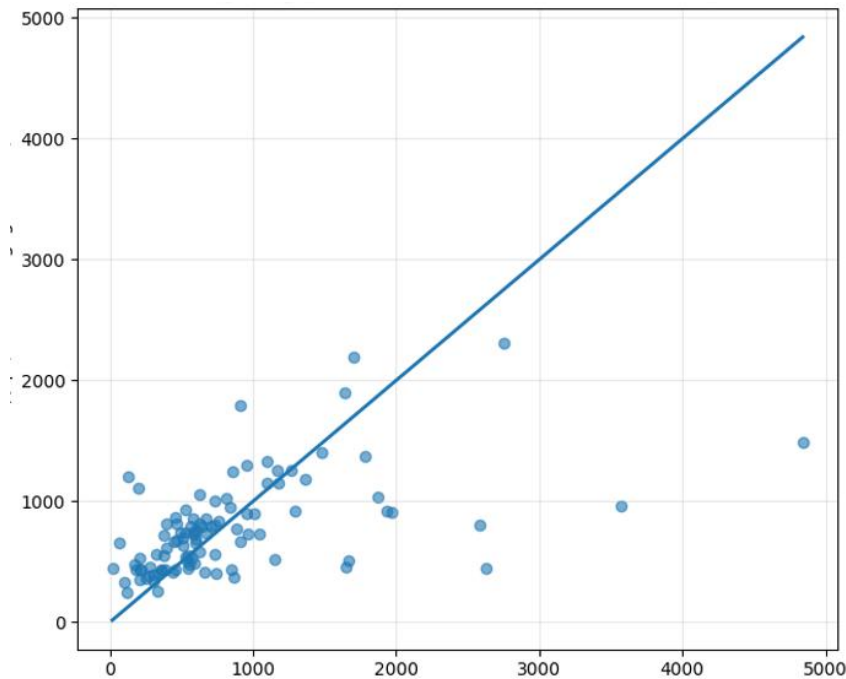


Figure 9. Comparison of actual (X-axis) and predicted (Y-axis) values of the Ridge model (log1p-back)

Note – compiled by the authors

This shows that the log1p transformation reduces the influence of outliers and increases the generalizability of the model. Figure 10 shows the residuals (y-axis) versus the predicted values (x-axis), illustrating the error structure of the Ridge (log1p-back) model.

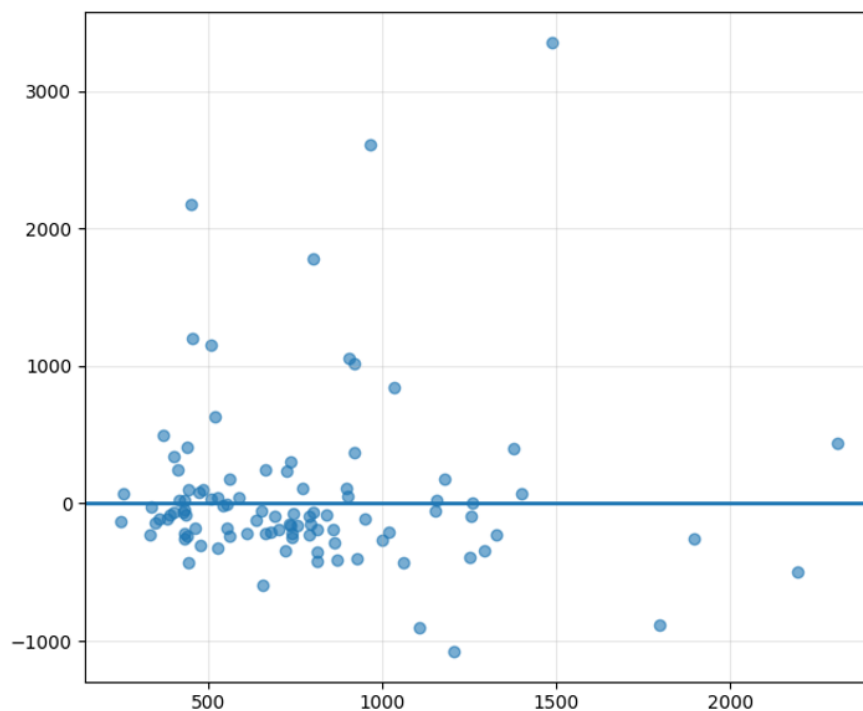


Figure 10. Distribution of residuals for the Ridge model (log1p-back)

Note – compiled by the authors

Compared to models using the original scale, using log1p provided more stable results in predicting performance. The results of the comparative evaluation of the models are presented in Table 4.

Table 4. Comparative results of the forecast quality of the models
in the scenario without leaks

No.	Model	Target variable	MAE	RMSE	R ²
1	Baseline (Dummy mean)	log1p-back	450.19	758.38	- 0.057
2	Ridge regression	raw	523.48	731.34	0.017
3	Ridge regression	log1p-back	357.05	629.66	0.271
4	Lasso regression	log1p-back	353.57	626.10	0.280
5	Random Forest	log1p-back	358.14	631.67	0.267
<i>Note – compiled by the authors</i>					

Note: log1p-back - the forecast results are transformed back to the original scale by applying a log1p transformation to the target variable.

Since the linear regression model is less stable than regularized models in the leak-free scenario and the performance is close to the baseline, the main discussion focuses on the results of Ridge/Lasso and Random Forest. It is worth noting that several machine learning models (linear regression, Ridge, Lasso and Random Forest) were considered during the study. The results show that the prediction performance of Ridge, Lasso and Random Forest models are close to each other in the log1p-back scenario (RMSE range is approximately 626–632). In this regard, ridge regression, which has high interpretability and good parametric stability, was chosen to formulate the main conclusions.

Since linear regression performed close to the baseline at the original scale, and Lasso and Random Forest models did not show statistically or practically significant advantages in terms of forecasting quality compared to ridge regression after log1p transformation, they were considered mainly as comparative and explanatory models in the results section. Even when changing the parameters of the Random Forest algorithm, no significant change in the RMSE was observed, indicating that the full potential of the ensemble model is not realized in the case of large data volumes and the aggregated nature of features. The results obtained showed that the use of complex and nonlinear models does not always provide an advantage in predicting indicators. In the scenario without data leaks, the limited data volume and the highly aggregated nature of the features prevented the random forest model from fully exploiting its potential. This situation once again confirms the importance of proper data preparation and feature selection, rather than increasing model complexity. In this sense, the relatively low R² values obtained in the no-leak scenario do not indicate a weakness of the models, but rather the true difficulty of the forecasting problem under consideration.

The Lifetime Engaged Users regression experiment and the intrusion detection task address two distinct machine-learning problems. The former evaluates the ability of regression models to predict continuous user-engagement outcomes, whereas the latter focuses on identifying malicious or anomalous behavior from security-related features. Therefore, the regression dataset is not treated as an IDS benchmark and its predictive performance is not interpreted as evidence of intrusion-detection effectiveness. The two experimental parts are connected at the methodological level through the broader study of machine-learning-based behavioral modeling.

CONCLUSION

The theoretical part provides a comprehensive overview of architectural solutions for firewalls, including both classic and modern NGFWs. The design features of the Palo Alto, pfSense, and MikroTik platforms are analyzed, and key functional differences are highlighted, such as the level of traffic inspection (from channel to application), filtering methods, support for the latest encryption protocols, and integration with intrusion detection systems. Particular attention is paid to the role of IPS, TLS decryption, sandbox analysis and machine learning used in NGFW.

During the experimental analysis, the issue of data leakage was taken into account, and known metrics were selected for training the model, which made it possible to create a scenario that excludes data leakage. The obtained results demonstrated the complexity of the task of forecasting the indicator in such a situation: linear models at the original scale did not demonstrate a significant advantage over the baseline. In this case, it was shown that the transformation of the target variable is more important than the choice of model. The experimental part of the study was implemented using Kali Linux as a generator of controlled malicious traffic. The test scenarios included SQL injection attacks, phishing, HTTP flooding, DDoS attacks using hping3, reverse shell simulations, downloading malicious binaries, and exploiting vulnerabilities using tools such as sqlmap, nmap, curl, wget, hydra, and others. Testing was performed on an isolated virtual infrastructure, including three environments with installed and configured NGFW Palo Alto, pfSense and MikroTik. For each attack type, logs were collected, security system activations were documented, and key metrics were calculated: the proportion of blocked threats, inspection depth, response speed, and support for modern security technologies.

To improve the score, a machine learning model in Python was applied using the Random Forest algorithm. Thus, the study confirmed the importance of a comprehensive approach to network protection, combining powerful NGFW filtering mechanisms with intelligent data analysis. Practical results confirm that the use of such solutions in corporate infrastructures ensures a high level of security, provided that policies are properly configured and regularly updated. The ability to detect attacks at the post-processing stage that are not recorded in real time was also demonstrated. The final findings of the study can serve as a basis for the creation of hybrid network security systems that combine advanced NGFWs and modern machine intelligence methods, which opens the way to the development of adaptive and scalable solutions for ensuring the security of corporate IT environments in the face of evolving cyber threats.

CONFLICT OF INTEREST: The authors declare no conflict of interest.

INSTITUTIONAL REVIEW BOARD STATEMENT: Not applicable.

INFORMED CONSENT STATEMENT: Not applicable.

DATA AVAILABILITY STATEMENT: Not applicable

ACKNOWLEDGEMENTS: The authors express their gratitude as to anonymous reviewers for valuable comments that helped improve the quality of the article.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES: The authors state that they did not use artificial intelligence (AI) tools in the preparation of this work.

REFERENCES

- Andariesta, D. T., & Wasesa, M. (2023). Machine learning models to predict the engagement level of Twitter posts: Indonesian e-commerce case study. *Procedia Computer Science*, 227, 823–832. <https://doi.org/10.1016/j.procs.2023.10.588>
- Ashrafi, N., Kuilboer, J.-P., Joshi, C., & Ran, W. (2022). Information security and compliance audit: A machine learning approach. *Government Information Quarterly*, 39(1), 126–131. <https://doi.org/10.5281/zenodo.17638989>

- Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25, 197–227. <https://doi.org/10.1007/s11749-016-0481-7>
- Bowei, Zh., Xiaoliang, W., Xie, R., Li, C., Huazheng, Zh., & Frank, J. (2023). A reputation mechanism based deep reinforcement learning and blockchain to suppress selfish node attack motivation in vehicular ad-hoc network. *Future Generation Computer Systems*, 139, 17–28. <https://doi.org/10.1016/j.future.2022.09.010>
- Davies, S. R., Wells, R., Zollo, F., & Roche, J. (2024). Unpacking social media “engagement”: A practice theory approach to science on social media. *Journal of Science Communication*, 23(6), Y02. <https://doi.org/10.22323/2.23060402>
- Dua, M., & Kunal. (2019). Machine learning approach to IDS: A comprehensive review. In *Proceedings of the 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA 2019)* (pp. 117–121). IEEE. <https://doi.org/10.1109/ICECA.2019.8822120>
- Eslami, S. P., Ghasemaghaei, M., & Hassanein, K. (2022). Understanding consumer engagement in social media: The role of product lifecycle. *Decision Support Systems*, 162, 113707. <https://doi.org/10.1016/j.dss.2021.113707>
- Gruschka, N., Jensen, M., & Iacono, L. (2023). Comparative analysis of risk assessment methods. *Lecture Notes in Computer Science*, 13745, 92–108. Springer. https://doi.org/10.1007/978-3-031-27324-7_8
- Hamid, Y., Sugumaran, M., & Balasaraswathi, V. R. (2016). IDS using machine learning: Current state of art and future directions. *British Journal of Applied Science & Technology*, 15(3), 1–22. <https://doi.org/10.9734/BJAST/2016/23668>
- Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2207.07048>
- Kumar, I., Bhatt, C., Patel, R., & others. (2020). Development of IDS using supervised machine learning. In *Soft Computing: Theories and Applications (Proceedings of SoCTA 2019)* (pp. 565–577). Springer Singapore. https://www.researchgate.net/profile/Chandradeep_Bhatt/-publication/342538220_Development_of_IDS_Using_Supervised_Machine_Learning
- Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences*, 9(20), 4396. <https://doi.org/10.3390/app9204396>
- Sudhir, V. (2020). Deep learning pipelines for financial compliance: Scalable document intelligence in regulated environments. *European Journal of Advances in Engineering and Technology*, 7(8), 126–131. <https://doi.org/10.5281/zenodo.17638989>
- Thakkar, A., & Lohiya, R. (2021). A review on machine learning and deep learning perspectives of intrusion detection systems for IoT: Recent updates, security issues, and challenges. *Archives of Computational Methods in Engineering*, 28(7), 3211–3243. <https://doi.org/10.1007/s11831-020-09496-0>
- Theodoridi, P. K., & Gkikas, D. C. (2025). Maximizing social media user engagement through predictive analytics in retail tourism: Identifying key performance indicators that trigger user interactions. *Applied Sciences*, 15(21), 11720. <https://doi.org/10.3390/app15211720>
- Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009). Intrusion detection by machine learning: A review. *Expert Systems with Applications*, 36(10), 11994–12000. <https://doi.org/10.1016/j.eswa.2009.05.029>

Авторлар туралы мәліметтер
Информация об авторах
Information about authors



Зиятбекова Гулзат Зиятбекқызы – PhD, қауымдастырылған профессор; әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан; Алматы технологиялық университеті, Алматы, Қазақстан; ҚР ҒЖБМ ҒК «Ақпараттық және есептеуіш технологиялар» институтының аға ғылыми қызметкері, Алматы қ., Қазақстан

Зиятбекова Гулзат Зиятбекқызы – PhD, ассоциированный профессор; Казахский национальный университет имени аль-Фараби, Алматы, Казахстан; Алматинский технологический университет, Алматы, Казахстан; старший научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, г. Алматы, Казахстан

Gulzat Ziyatbekova – PhD, Associate Professor; Al-Farabi Kazakh National University, Almaty, Kazakhstan; Almaty Technological University, Almaty, Kazakhstan; Senior Researcher at the Institute of Information and Computational Technologies SC MSHE RK, Almaty, Kazakhstan

e-mail: ziyatbekova1@gmail.com

ORCID: <https://orcid.org/0000-0002-9290-6074>



Кыдырбаева Галия Турыспаевна – педагогика ғылымдарының кандидаты, I. Жансүгіров атындағы Жетісу университеті, Талдықорған қ., Қазақстан

Кыдырбаева Галия Турыспаевна – кандидат педагогических наук, Жетысуский университет имени И. Жансугурова, г. Талдыкорган, Казахстан

Galiya Kydyrbayeva – Candidate of Pedagogical Sciences, Zhetysu university named after I. Zhansugurov, Taldykorgan, Kazakhstan

e-mail: kidirbaeva@gmail.com

ORCID: <https://orcid.org/0000-0002-3050-1688>



Рахимжанова Гульмира Каирбековна – информатика және математика бойынша маман-әдіскер, «Дарын» оқу-әдістемелік орталығы, Талдықорған қ., Қазақстан

Рахимжанова Гульмира Каирбековна – специалист-методолог по информатике и математике, учебно-методологический центр «Дарын», г. Талдыкорган, Казахстан

Gulmira Rahimzhanova – specialist-methodologist in computer science and mathematics, educational and methodological center "Daryn", Taldykorgan, Kazakhstan

e-mail: rahimzhanovag@mail.ru

ORCID: <https://orcid.org/0009-0001-6571-4560>



Шаметова Гауһар Құттымұратқызы – әл-Фараби атындағы Қазақ ұлттық университетінің «Киберқауіпсіздік және криптология» кафедрасының докторанты, Алматы қ., Қазақстан
Шаметова Гауһар Құттымұратқызы – докторант кафедрасы «Кибербезопасность и криптология» Казахского национального университета имени аль-Фараби, г. Алматы, Казахстан

Gauhar Shametova – Doctoral student of the Department of Cybersecurity and Cryptology, Al-Farabi Kazakh National University, Almaty, Kazakhstan

e-mail: gauharshametova@gmail.com

ORCID: <https://orcid.org/0009-0001-4099-0112>



Молдакалыкова Баян Жоямергеновна – Тұран Университеті Ақпараттық технологиялар жоғары мектебінің сениор-лекторы, Алматы қ., Қазақстан

Молдакалыкова Баян Жоямергеновна – Сениор-лектор Высшей школы информационных технологий университета Туран, г. Алматы, Казахстан

Bayan Moldakalykova – Senior Lecturer of the Higher School of Information Technology at Turan University, Almaty, Kazakhstan,

e-mail: b.moldakalykova@turan-edu.kz

ORCID: <https://orcid.org/0000-0003-4840-1228>



Білімжанқызы Назерке – білім беру магистры, Тұран Университетінің ассистент-лекторы, Алматы қ., Қазақстан

Білімжанқызы Назерке – магистр образования, ассистент-лектор университета Туран, г. Алматы, Казахстан

Bilimzhankyzy Nazerke – Master of Education, Assistant Lecturer at Turan University, Almaty, Kazakhstan,

e-mail: nazerkebilimzhankyzy@gmail.com

ORCID: <https://orcid.org/0009-0008-2081-9810>